Understanding the new Language Tags

Richard Ishida
W3C Internationalization Activity Lead

Objectives

- understand the basic syntax of the replacement for RFC 3066
- understand how to use RFC 3066bis language tags
- be familiar with the language registry
- be aware of future enhancements to RFC 3066bis due in the near term

After reading this you should:

Notes

Based on an article by Addison Phillips at

http://www.w3.org/International/articles/bcp47/

Overview of syntax

Language tags

xml:lang="…"

etc.

Overview of syntax

RFC 3066

language – script
IANA registered tag

- ISO 639 language codes
- ISO 3166 country codes

en
en-GB
en-scouse
Subtags

- Subtags are 1 – 8 characters long, but length is significant in some cases.
- Subtags restricted to a-z, A-Z, 0-9.
- Case not significant.
- All subtags in one (new) registry.
- All subtags, except grandfathered, are generative.
- 34 grandfathered tags (8 already obsolete, 10 heading for obsoletion).

http://www.iana.org/assignments/language-subtag-registry

Primary language tags

- Always required.
- ISO 639 code or registered value.
- Codes available only from new IANA registry.
- Two-letter codes provided if available, 3-letter if not.

Script tags

- ISO 15924 code.
- Only one, directly after language.
- 4 letters long.
- Use only if needed.
- ‘Suppress-Script’ labels in registry, eg. en-Latn.

Variant tags

- Individually registered values.
- Indicates dialects or script variations not covered by lang+region.
- Registry fields indicate appropriate usage, eg. "Prefix: sl"
- Script and region codes not required.

Region tags

- ISO 3166-1 code or UN M.49 region code.
- 2 letter alpha or 3 digit codes.
- Only one following language and any script codes.
- Script code not required.

Addison Phillips

"... avoid script tags except where they add useful distinguishing information."

"For virtually any content that does not use a script tag today, it remains the best practice not to use one in the future."

Addison Phillips
Extensions and private use

- allows for addition of future extensions to language tags
- introduced by a 'singleton'
- private use singleton is 'x'
- no extensions currently registered

Subtag registry

- **Basic fields**
  - **Type**: language
  - **Subtag**: cs
  - **Description**: Czech
  - **Added**: 2005-10-16
  - **Suppress-Script**: Latn
  - **Type**: language
  - **Subtag**: cu
  - **Description**: Church Slavic
  - **Description**: Old Slavonic
  - **Description**: Church Slavonic
  - **Description**: Old Bulgarian
  - **Description**: Old Church Slavonic
  - **Added**: 2005-10-16
  
- **Type**: language
  - **Subtag**: cv
  - **Description**: Chuvash
  - **Added**: 2005-10-16

- **Deprecated tags**
  - **Type**: region
  - **Subtag**: TP
  - **Description**: East Timor
  - **Added**: 2005-10-16
  - **Preferred-Value**: TL
  - **Deprecated**: 2002-11-15
  - **Type**: grandfathered
  - **Tag**: i-navajo
  - **Description**: Navajo
  - **Added**: 1997-09-19
  - **Preferred-Value**: nv
  - **Deprecated**: 2000-02-18
  - **Comments**: replaced by ISO code nv

Remaining work

- Last Call, Publication
- Matching
- Naming (what is BCP 477?)
- ISO 639-3 and Macro Languages

http://www.w3.org/International/core/langtags/rfc3066bis.html
Extended language subtags

- extended_language
  - zh-cmn
  - zh-cmn-Hans
  - zh-cmn-Hant-HK
  - zh-gan
  - zh-hak
  - zh-yue-Hant-HK

Refine 'macro-language' codes such as zh based on ISO 639-3 already allowed for in RFC 3066bis. Usage limited to specified macro-languages.

Further reading

Spec locations
http://www.w3.org/International/core/language/rfc3066bis.html
Overview by Addison Phillips
http://www.w3.org/International/articles/bcp47/